

# 习题一

151250104, 卢以宁

2017 年 3 月 12 日

## Problem 1

若数据包含噪声, 则假设空间中有可能不存在与所有训练样本都一致的假设, 此时的版本空间是什么? 在此情形下, 试设计一种归纳偏好用于假设选择。

**Solution.** 此时的版本空间为空。选择的假设应符合尽量多的样本; 同时可遵循奥卡姆剃刀原则, 通过正则项等方法避免过拟合。

## Problem 2

对于有限样例, 请证明

$$\text{AUC} = \frac{1}{m^+m^-} \sum_{x^+ \in D^+} \sum_{x^- \in D^-} \left( \mathbb{I}(f(x^+) > f(x^-)) + \frac{1}{2} \mathbb{I}(f(x^+) = f(x^-)) \right)$$

**Proof.** 由定义,

$$\text{AUC} = \frac{1}{2} \sum_{i=1}^{m-1} (x_{i+1} - x_i)(y_i + y_{i+1})$$

而设排序后第  $k$  个样例为  $s_k, (k = 2, 3, 4, \dots, m)$ , 对应的坐标为  $(x_k, y_k)$ , 则

$$(x_k - x_{k-1})(y_k + y_{k-1}) = \begin{cases} 0 & s_k \in D^+ \\ \frac{1}{m^-} \cdot y_k & s_i \in D^- \end{cases}$$

其中,  $y_k$  是由  $s_k$  及其之前的真正例数决定的。

$$y_k = \frac{1}{m^+} \cdot \sum_{x^+ \in D^+} \left( \mathbb{I}(f(x^+) > f(x_k)) + \frac{1}{2} \mathbb{I}(f(x^+) = f(x_k)) \right)$$

其中,  $\frac{1}{2}$  表示当某个真正例的预测值和  $f(x_k)$  相等时, 它有一半的几率出现在  $s_i$  之前。综上所述可得,

$$\text{AUC} = \frac{1}{m^+m^-} \sum_{x^+ \in D^+} \sum_{x^- \in D^-} \left( \mathbb{I}(f(x^+) > f(x^-)) + \frac{1}{2} \mathbb{I}(f(x^+) = f(x^-)) \right)$$

□

### Problem 3

在某个西瓜分类任务的验证集中，共有 10 个示例，其中有 3 个类别标记为“1”，表示该示例是好瓜；有 7 个类别标记为“0”，表示该示例不是好瓜。由于学习方法能力有限，我们只能产生在验证集上精度 (accuracy) 为 0.8 的分类器。

(a) 如果想要在验证集上得到最佳查准率 (precision)，该分类器应该作出何种预测？

此时的查全率 (recall) 和 F1 分别是多少？

(b) 如果想要在验证集上得到最佳查全率 (recall)，该分类器应该作出何种预测？

此时的查准率 (precision) 和 F1 分别是多少？

**Solution.** 如下：

(a) 因为准确率是 0.8 所以分错 2 个。由于有 3 个正例，至少有一个 TP，标记它为正例，其他标记为反例，accuracy 为 1

$$R = \frac{1}{3}, F_1 = \frac{2 \times P \times R}{P + R} = \frac{1}{2}$$

(b) 将所有正例预测为正例，查全率为 100%。此时将 2 个坏瓜误认为好瓜。

$$P = \frac{TP}{TP + FP} = \frac{3}{5}, F_1 = \frac{2 \times P \times R}{P + R} = \frac{3}{4}$$

### Problem 4

在数据集  $D_1, D_2, D_3, D_4, D_5$  运行了  $A, B, C, D, E$  五种算法，算法比较序值表如表 1 所示：

表 1: 算法比较序值表

数据集	算法 A	算法 B	算法 C	算法 D	算法 E
$D_1$	2	3	1	5	4
$D_2$	5	4	2	3	1
$D_3$	4	5	1	2	3
$D_4$	2	3	1	5	4
$D_5$	3	4	1	5	2
平均序值	3.2	3.8	1.2	4	2.8

使用 Friedman 检验 ( $\alpha = 0.05$ ) 判断这些算法是否性能都相同。若不相同，进行 Nemenyi 后续检验 ( $\alpha = 0.05$ )，并说明性能最好的算法与哪些算法有显著差别。

**Solution.**

$$\begin{aligned}
 T_{\chi^2} &= \frac{k-1}{k} \cdot \frac{12N}{k^2-1} \cdot \sum_{i=1}^k \left(r_i - \frac{k+1}{2}\right)^2 \\
 &= \frac{12N}{k(k+1)} \cdot \left(\sum_{i=1}^k r_i^2 - \frac{k(k+1)^2}{4}\right) \\
 &= \frac{60}{5 \cdot 6} \cdot \left(\sum_{i=1}^5 r_i^2 - \frac{5 \cdot 36}{4}\right) \\
 &= 9.92 \\
 T_F &= \frac{(N-1)T_{\chi^2}}{N(k-1) - T_{\chi^2}} = 3.93
 \end{aligned}$$

当  $\alpha = 0.05, N = 5, k = 5$  时, F 检验的临界值为 3.007, 故拒绝假设, 认为算法性能不相同。然后使用 Nemenyi 后续检验得到

$$CD = q_\alpha \cdot \sqrt{\frac{k(k+1)}{6N}} = 2.728 \cdot \sqrt{\frac{5 \cdot 6}{6 \cdot 5}} = 2.728$$

经过比较得到, 算法 C 与算法 D 的性能显著不同, 其余算法之间无显著不同。